Conditional inferential models: combining information for prior-free probabilistic inference

Ryan Martin
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
rgmartin@uic.edu

Chuanhai Liu
Department of Statistics
Purdue University
chuanhai@purdue.edu

November 8, 2012

Abstract

The inferential model (IM) framework provides valid prior-free probabilistic inference by focusing on predicting unobserved auxiliary variables. But efficient IM-based inference can be challenging when this auxiliary variable is high-dimensional. Here we show that characteristics of the auxiliary variable are often fully observed and, in such cases, a simultaneous dimension reduction and information aggregation can be achieved by conditioning. This proposed conditioning strategy leads to efficient IM inference, and casts new light on Fisher's notions of sufficiency, conditional inference, and also Bayesian inference. A differential equation-driven selection of a conditional association is developed, and we prove a conditional IM validity theorem under some conditions. Some problems, however, may not admit a valid conditional IM of the standard form. For such cases, we propose a more flexible class of conditional IMs based on localization. The take-away message is that the conditional IM framework developed herein provides valid and efficient prior-free probabilistic inference in a variety of challenging problems.

Keywords and phrases: Ancillary; auxiliary variable; Bayes; belief function; differential equation; sufficiency; predictive random set; validity.

1 Introduction

Fisher's brand of statistical inference (Fisher 1973) is often viewed as a middle-ground between the Bayesian and frequentist approaches. Two important examples are his fiducial argument and his ideas on conditional inference. Perhaps influenced by Fisher's ideas, a current focus in foundational research is on achieving some kind of compromise between

the Bayesian and frequentist ideals. See, for example, recent work on fiducial inference (Hannig 2009, 2012; Hannig and Lee 2009), confidence distributions (Xie and Singh 2012; Xie et al. 2011), Dempster–Shafer theory (Dempster 2008; Shafer 2011), and objective Bayes with default, reference, and/or data-dependent priors (Berger 2006; Berger et al. 2009; Fraser 2011; Fraser et al. 2010). Recently Martin and Liu (2012a) have laid out the details of a promising new *inferential model* (IM) approach; see, also, Martin et al. (2010) and Zhang and Liu (2011). IMs take the usual input—sampling model and observed data—and produce prior-free, posterior-probabilistic measures of certainty about any assertion/hypothesis of interest, with an almost automatic calibration property. The fundamental idea is that uncertainty about the parameter of interest θ, given observed data $X = x$, is fully characterized by the unobserved value $u^\star$ of an associated auxiliary variable U. So the problem of inference about θ can be translated into one of predicting this unobserved value $u^\star$ with a predictive random set. In Section 2 we briefly review the construction and basic properties of IMs.

The discussion in Martin and Liu (2012a) focuses on the case where θ and $u^\star$ are of the same dimension. But there are many problems, e.g., iid data from scalar parameter models, where the dimension of the auxiliary variable is much greater than that of the parameter. In such cases, efficiency can be gained by first reducing the dimension of the auxiliary variable to be predicted, though it is not at all obvious how to perform this dimension reduction in general. In this paper we focus our attention on an auxiliary variable dimension reduction step based on conditioning. The critical observation here is that, typically, certain functions of the auxiliary variables are fully observed. So, by conditioning on those observed characteristics of the auxiliary variable, we can effectively reduce the dimension of the unobserved characteristics to be predicted. The fundamental result, proved in Section 3.2, is that this reduction is accomplished without loss of information. Therefore, we can view this dimension-reduction approach as a tool for combining information about θ across samples—a counterpart to Bayes' theorem and Fisher's sufficiency. From the resulting lower-dimensional auxiliary variable representation, we proceed to construct what is called a *conditional IM*. In Section 3.4, we give a general validity theorem that establishes a desirable calibration property of the conditional IM, which helps facilitate a common interpretation across users and experiments.

Finding the dimension-reduced representation, the subject of Section 4, is sometimes a familiar task. For example, when the minimal sufficient statistic has dimension matching that of the parameter, the conditional IM is exactly that obtained by working directly with said statistic. In other cases, finding the lower-dimensional representation is not so simple, analogous to finding ancillary statistics in the classical context. For this, we propose a new differential equation-driven technique for identifying observed characteristics of the auxiliary variable. Two classical conditional inference problems are worked out in Section 5, one showing how the proposed differential equation technique leads to an additional dimension reduction beyond what ordinary sufficiency provides. So, besides the development of conditional IMs, the proposed framework also casts new light on the familiar notion of sufficiency, as well as Fisher's attractive but elusive ideas on ancillary statistics, conditional inference, etc.

In some cases, however, it may not be possible to produce a valid conditional IM with these somewhat standard techniques. For this, in Section 6, we propose an extension of the conditional IM framework which allows the lower-dimensional auxiliary variable

representation to depend on θ in a certain sense. We refer to these as *local conditional IMs*, and we describe their construction and prove a validity theorem. An important example of such a problem is the bivariate normal model with known means and variances but unknown correlation. For this example, we construct a local conditional IM based on a modification of the differential equations technique, and provide the results of a simulation study that shows that our conditional plausibility intervals outperform the classical $r^\star$-driven asymptotically approximate confidence intervals (Barndorff-Nielsen 1986; Fraser 1990) in both small and large samples. A local conditional IM analysis of the variance-components problem in Cox (1980) is also given.

We conclude in Section 7 with some remarks. In particular, we highlight the main contributions of this paper and discuss some potential extensions of these important ideas and results, including a related IM strategy for nuisance parameter problems.

2 Review of IMs

2.1 Notation and construction

To fix notation, let X be the observable data, taking values in a space $\mathbb{X}$, and let θ be the parameter of interest, taking values in the parameter space Θ. The starting point of the IM framework is similar to that of fiducial, in the sense that an auxiliary variable, denoted by U and taking values in a space $\mathbb{U}$ with probability measure P_U, is associated with X and θ. It is this association, together with the distribution $U \sim \mathsf{P}_U$, that characterizes the sampling distribution $X \sim \mathsf{P}_{X|\theta}$. In particular, if we write this association as

$$X = a(\theta, U), \quad U \sim \mathsf{P}_U, \tag{2.1}$$

then we require that X generated according to the above "algorithm," i.e., first sample $U \sim \mathsf{P}_U$ and set $X = a(\theta, U)$ for given θ, have distribution $\mathsf{P}_{X|\theta}$.

Compared to fiducial inference, which employs the sampling distribution P_U *after* $X = x$ is observed, the IM approach takes a different perspective. Specifically, the IM approach treats the unobserved value $u^\star$ of U, which is tied to the observed data $X = x$ and the *true value* of θ, as the fundamental quantity. Then the goal is to predict this unobserved value $u^\star$ with a random set before conditioning on $X = x$ and inverting (2.1).

Here we follow Martin and Liu (2012a); see their paper for full details. Start with a collection $\mathbb{S}$ of P_U-measurable subsets of $\mathbb{U}$, assumed to contain $\varnothing$ and $\mathbb{U}$. This collection will serve as the support of the predictive random set. For optimal predictive random sets, it suffices to assume that the collection $\mathbb{S}$ is nested, i.e., either $S \subseteq S'$ or $S' \subseteq S$ for all $S, S' \in \mathbb{S}$. We can now define the predictive random set $\mathcal{S}$, supported on $\mathbb{S}$, with distribution $\mathsf{P}_{\mathcal{S}}$ satisfying

$$\mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq K\} = \sup_{S \in \mathbb{S}: S \subseteq K} \mathsf{P}_U(S), \quad K \subseteq \mathbb{U}.$$

$\mathsf{P}_{\mathcal{S}}\{\mathcal{S} \subseteq \cdot\}$ is like the "distribution function" of the random set $\mathcal{S}$. Predictive random sets constructed in this way are called *admissible*. In scalar θ problems, P_U is often $\mathsf{Unif}(0,1)$, so an important example of an admissible predictive random set is

$$\mathcal{S} = \{u : |u - 0.5| \leq |U - 0.5|\}, \quad U \sim \mathsf{Unif}(0,1). \tag{2.2}$$

Martin and Liu (2012a, Corollary 1) show that this $\mathcal{S}$ has a variety of good properties, and these good properties often carry over to the corresponding IM, provided that the set $\Theta_x(u) = \{\theta : x = a(\theta, u)\}$ moves monotonically[1] as a (set-valued) function of u for each x. We shall also employ this "default" predictive random set in our examples herein.

The following three steps—association, predict, and combine—described in Martin and Liu (2012a), together define an IM.

A-step. Associate X, θ, and $U \sim \mathsf{P}_U$, consistent with the sampling distribution $X \sim \mathsf{P}_{X|\theta}$, such that, for all (x, u), there is a unique subset $\Theta_x(u) = \{\theta : x = a(\theta, u)\} \subseteq \Theta$, possibly empty, containing all possible candidate values of θ given (x, u).

P-step. Predict the unobserved value $u^\star$ of U associated with the observed data by an admissible predictive random set $\mathcal{S}$.

C-step. Combine $\mathcal{S}$ and the association $\Theta_x(u)$ specified in the A-step to obtain

$$\Theta_x(\mathcal{S}) = \bigcup_{u \in \mathcal{S}} \Theta_x(u). \tag{2.3}$$

Then compute the *belief function*

$$\mathsf{bel}_x(A; \mathcal{S}) = \mathsf{P}_{\mathcal{S}}\{\Theta_x(\mathcal{S}) \subseteq A \mid \Theta_x(\mathcal{S}) \neq \varnothing\}, \tag{2.4}$$

where $A \subseteq \Theta$ is the assertion/hypothesis about θ of interest.

The belief function is just one part of the inferential output. Since the belief function is sub-additive, i.e., $\mathsf{bel}_x(A; \mathcal{S}) + \mathsf{bel}_x(A^c; \mathcal{S}) \leq 1$, one actually needs both $\mathsf{bel}_x(A; \mathcal{S})$ and $\mathsf{bel}_x(A^c; \mathcal{S})$ to summarize the information in x concerning the truthfulness of assertion A. In some cases, it is more convenient to report the *plausibility function*

$$\mathsf{pl}_x(A; \mathcal{S}) = 1 - \mathsf{bel}_x(A^c; \mathcal{S}). \tag{2.5}$$

Then the pair $(\mathsf{bel}_x, \mathsf{pl}_x)(A; \mathcal{S})$ characterize the IM output. Note that there are reasons one might consider using a different predictive random set for each of $\mathsf{bel}_x(A; \cdot)$ and $\mathsf{bel}_x(A^c; \cdot)$; see Martin and Liu (2012a) and Martin et al. (2012). These two papers also provide a variety of examples illustrating the construction of IMs.

Without practical loss of generality, assume that $\{\mathsf{P}_{X|\theta} : \theta \in \Theta\}$ has a common dominating measure, say μ. Then we require that $\mathsf{bel}_x(A; \mathcal{S})$ be a μ-measurable function in x for all A. This is easy to check in examples, but general sufficient conditions are more elusive. To keep presentation simple, we shall mostly ignore these technical concerns.

2.2 Validity of IMs

The performance of a predictive random set is measured through the sampling behavior of the corresponding belief function, as a function of $X \sim \mathsf{P}_{X|\theta}$, for a given assertion A. Given $\mathcal{S}$, the corresponding IM is *valid for* A if the belief function satisfies

$$\sup_{\theta \notin A} \mathsf{P}_\theta\big\{\mathsf{bel}_X(A; \mathcal{S}) \geq 1 - \alpha\big\} \leq \alpha, \quad \forall\, \alpha \in (0, 1). \tag{2.6}$$

[1] A set-valued mapping $\Theta_x(\cdot)$ moves monotonically right (left) if, for $u_2 > u_1$, for any $\theta \in \Theta_x(u_1)$, there exists a $\theta' \in \Theta_x(u_2)$ such that $\theta' \geq \theta$ ($\theta' \leq \theta$).

The IM is simply called *valid* if it is valid for all A. In other words, the IM is valid for A if $\mathsf{bel}_X(A;\mathcal{S})$ is stochastically no larger than $\mathsf{Unif}(0,1)$ when $X \sim \mathsf{P}_{X|\theta}$ with $\theta \notin A$. That is, if A is *false*, then the amount of support in data X for A will be large only for a relatively small proportion of X values. Martin and Liu (2012a, Theorem 1) show that this validity property is easy to arrange: it holds for all A whenever the predictive random set $\mathcal{S}$ is admissible in the sense described above. In this case, if the IM is valid for all A, then (2.6) can be equivalently stated in terms of the plausibility function:

$$\sup_{\theta \in A} \mathsf{P}_{X|\theta}\{\mathsf{pl}_X(A;\mathcal{S}) \leq \alpha\} \leq \alpha, \quad \forall\, \alpha \in (0,1). \tag{2.7}$$

This formulation is occasionally more convenient than (2.6).

There are two important consequences of the validity theorem. First, it helps determine an objective scale on which the belief probabilities can be interpreted. Martin et al. (2012) discuss notions of meaningfulness *within* and *across* experiments, and they argue that the validity theorem is critical to having both these interpretations simultaneously. Bayesian, fiducial, and Dempster–Shafer probabilities are subjective and, therefore, unlike valid IMs, they do not have a common scale on which they can be interpreted. Second, if one so chooses, the validity theorem allows one to use the IM output to construct frequentist decision procedures with control on error rates. For example, one can construct a $100(1-\alpha)\%$ plausibility region for θ:

$$\{\theta : \mathsf{pl}_x(\theta;\mathcal{S}) > \alpha\}. \tag{2.8}$$

It follows easily from (2.7) that this plausibility region has nominal $1-\alpha$ coverage probability. But we should emphasize here that, although plausibility functions can be used to construct frequentist procedures, the interpretation is quite different. For example, the plausibility region is understood as the collection of points such that each is *individually* sufficiently plausible, given $X = x$. Confidence/credible regions, on the other hand, do not have such a simple yet sharp interpretation.

3 Conditional IMs

3.1 Motivation

Most of the examples in Martin and Liu (2012a) have a scalar auxiliary variable U. This makes construction of efficient predictive random sets relatively easy. However, scalar auxiliary variables is an extremely special case. To see this, suppose $X_1, \ldots, X_n$, $n > 1$, are independent $\mathsf{N}(\theta, 1)$ observations with common unknown mean θ. In vector notation, an association is $X = \theta 1_n + U$, where 1_n is an n-vector of unity, and $U \sim \mathsf{N}_n(0, I)$. Without careful thinking, it seems that one must predict an n-dimensional auxiliary variable $u^\star$. But efficient prediction of $u^\star$ would be challenging if n is even moderately large, so reducing the dimension of $u^\star$—ideally to one dimension—would be a desirable first step. In a classical framework, one can avoid this dimensionality difficulty by reducing X to a sufficient statistic, e.g., $\overline{X}$, and construct an association from there. In this and the next section, we develop an IM-based theory that justifies this sort of intuition.

3.2 Dimension reduction via conditioning

As indicated in Section 3.1, efficient prediction of the unobserved auxiliary variable can be difficult in moderate- to high-dimensions. Here we investigate an approach by which a simultaneous aggregation of information and auxiliary variable dimension reduction can be achieved. The intuition is that some functions of $u^\star$ are actually observed, so these characteristics do not need to be predicted. This provides a sort of aggregation of information. Furthermore, these observed characteristics also provide a dimension reduction which can help to better predict those aspects that remain unobserved. The general strategy is as follows:

- Identify an observed characteristic of the auxiliary variable whose distribution is free (or at least mostly free) of θ, and
- define a conditional association that relates a $\dim(\Theta)$-dimensional function of the auxiliary variable to θ and some function of observable data X.

The second step is familiar, as it relates to working with, e.g., minimal sufficient statistics of $\dim(\Theta)$ dimension. The first step, however, is less familiar and can be difficult; see Section 4. In Theorem 1 below, we show that this auxiliary variable dimension-reduction scheme can be accomplished without loss of information. "Information" has a precise meaning in the classical setting, via likelihood. Our framework is different, so we must first explain what is meant by "without loss of information" in this setting.

In the theorem that follows, a *naive IM* is that based on a singleton predictive random set. For example, in the baseline association, the naive IM uses $\mathcal{S} = \{U\}$ with $U \sim \mathsf{P}_U$. This is a poor choice of predictive random set from a practical point of view, but it is a convenient choice when the goal is to compare two candidate associations, especially when the auxiliary variable spaces are quite different. Given observation $X = x$, the triplet (x, a, P_U) is all that is needed to evaluate the naive IM's belief function.

Theorem 1. *Suppose that the relationship $x = a(u, \theta)$ in the baseline association (2.1) can be decomposed into the system:*

$$H(x) = \psi_H(u), \tag{3.1a}$$
$$T(x) = a_T(\psi_T(u), \theta), \tag{3.1b}$$

where $x \mapsto (T(x), H(x))$ and $u \mapsto (\psi_T(u), \psi_H(u))$ are one-to-one and free of θ. Let $(V_T, V_H) \in \mathbb{V}_T \times \mathbb{V}_H$ be the image of U under (ψ_T, ψ_H), and let $\mathsf{P}_{V_T|h}$ be a version of the conditional distribution of V_T, given $V_H = h$, $h \in H(\mathbb{X})$. Then the baseline association (2.1) and that determined by $T(x) = a_T(v_T, \theta)$ and $\mathsf{P}_{V_T|H(x)}$ are equivalent for inference on θ in the sense that, for any observed $X = x$, the triplets (x, a, P_U) and $(T(x), a_T, \mathsf{P}_{V_T|H(x)})$ produce identical naive IM belief functions.

The decomposition (3.1) boils down to a specification of a particular hierarchical representation of the sampling model for X. Indeed, for functions H and T as in the theorem, with $V_H = \psi_H(U)$, and $V_T = \psi_T(U)$, data $X \sim \mathsf{P}_{X|\theta}$ can be simulated as follows.

1. Sample (V_T, V_H) by sampling $V_H \sim P_{V_H}$ and $V_T \mid V_H \sim \mathsf{P}_{V_T|V_H}$;
2. Obtain X by solving the system $H(X) = V_H$ and $T(X) = a_T(V_T, \theta)$.

This hierarchical model representation provides the following insight: when $X = x$ is observed, so too is the value of V_H, and this knowledge can be used to update the auxiliary variable distribution, analogous to Bayes' theorem. Hence, Theorem 1 provides a sort of aggregation of information. Further extensions and consequences of Theorem 1 are collected in a series of remarks in Section 3.3.

Another important consequence of Theorem 1 is as follows. Let $V_T = \psi_T(U)$, taking values in $\mathbb{V}_T = \psi_T(\mathbb{U})$. Then the theorem states that it suffices to consider a new conditional association that connects transformed data $T(x)$, transformed auxiliary variable v_T, and parameter θ through the mapping

$$T(x) = a_T(v_T, \theta), \quad T(x) \in T(\mathbb{X}), \quad v_T \in \mathbb{V}_T, \tag{3.2}$$

which seems to ignore the first equation (3.1a). However, on the contrary, (3.1a) can never be ignored, because it is used to update the auxiliary variable distribution from P_U to $\mathsf{P}_{V_T|H(x)}$. If it happens that V_T and V_H are independent, then the first constraint (3.1a) actually plays no role. The important point is not the simple recasting of the association that is important; rather, it is that ψ_T can often be chosen so the new auxiliary variable V_T is of lower dimension than U. In fact, V_T will often have dimension the same as that of θ. In addition to providing a sort of summary of the data, like in the classical context, this auxiliary variable dimension reduction has a unique advantage in the IM context: efficient predictive random sets for the lower-dimensional V_T are easier to construct.

Once a decomposition (3.1) is available, construction of a new conditional IM follows exactly as in Section 1. To simplify the presentation later on, here we restate the simple three-step construction of a *conditional IM*.

A-step. Associate $T(x)$ and θ with the new auxiliary variable $v_T = \psi_T(u)$ to get the collection of sets $\Theta_{T(x)}(v_T) = \{\theta : T(x) = a_T(v_T, \theta)\}$, $v_T \in \mathbb{V}_T$, based on (3.2).

P-step. Fix $h = H(x)$. Predict the unobserved value $v_T^\star$ of V_T with a *conditionally admissible* predictive random set $\mathcal{S} \sim \mathsf{P}_{\mathcal{S}|h}$ (see Section 3.4).

C-step. Combine results of the A- and P-steps to get

$$\Theta_{T(x)}(\mathcal{S}) = \bigcup_{v_T \in \mathcal{S}} \Theta_{T(x)}(v_T) \subseteq \Theta. \tag{3.3}$$

Then the corresponding conditional belief and plausibility functions are given by

$$\begin{aligned} \mathsf{bel}_{T(x)|h}(A; \mathcal{S}) &= \mathsf{P}_{\mathcal{S}|h}\{\Theta_{T(x)}(\mathcal{S}) \subseteq A \mid \Theta_{T(x)}(\mathcal{S}) \neq \varnothing\} \\ \mathsf{pl}_{T(x)|h}(A; \mathcal{S}) &= 1 - \mathsf{bel}_{T(x)|h}(A^c; \mathcal{S}). \end{aligned} \tag{3.4}$$

These functions can be used for inference on θ just like those in Section 2.

3.3 Remarks

Remark 1. Theorem 1 holds for more general decompositions (3.1). That is, one may replace "$H(x) = \psi_H(u)$" in (3.1a) with "$c(x, u) = 0$" for a function c. However, this more general "non-separable" case does not fit into the context of the conditional validity theorem; see Theorem 2. So although (3.1a) is not necessary for Theorem 1, the more general version is dangerous since the corresponding IM may not have the proper calibration properties. We will have more to say about this in Section 6.

Remark 2. In some cases, it will be convenient to rewrite the conditional association (3.2) to absorb the dependence on $h = H(x)$ into the function a_T, so that the auxiliary variable and predictive random set can have a fixed distribution, independent of observed x. That is, (3.2) may be rewritten as

$$T(X) = \tilde{a}_T(W, \theta, h), \quad W \sim \mathsf{P}_W,$$

where P_W is free of both θ and h. For example, if V_T is one-dimensional, let F_h be its distribution function. If this distribution is continuous, then $W = F_h(V_T) \sim \mathsf{Unif}(0,1)$; in that case, set $\tilde{a}_T(w, \theta, h) = a_T(F_h^{-1}(w), \theta)$. See Section 5.

Remark 3. An immediate consequence of Theorem 1 is that, if two different decompositions of the form (3.1) are available, then the two corresponding naive conditional IMs are equivalent in the sense that their belief functions are identical. Therefore, although different baseline associations could be chosen, and a variety of different ways to construct a decomposition (3.1) are available, there is still a notion of conditional IM uniqueness: the belief functions based on singleton predictive random sets are identical. This equivalence can be extended beyond the singleton predictive random set case whenever one conditional IM is a one-to-one reparametrization of the other. Basically, our type of conditional IM uniqueness stems from the two-step hierarchical representation of the sampling model given above. That is, any decomposition (3.1) specifies a hierarchical representation, and since these are all equivalent, so too are the corresponding conditional IMs, modulo choice of predictive random set.

Remark 4. There are clearly some close connections between the result in Theorem 1 and Fisher's notion of sufficiency. At a very high level, both theories provide a sort of dimension reduction. The key difference between the two is that sufficiency focuses on reducing the dimension of the observable data, while Theorem 1 focuses on reducing the dimension of the unobservable auxiliary variable. Although the conditional IM can, in some cases, correspond to a sufficient statistic-type of reduction but, in light of the equivalence in Remark 3, this is not necessary. In this sense, sufficiency, and the related notions of completeness, minimal sufficiency, etc, are not fundamental concepts in the IM framework. Moreover, in the classical framework, conditional inference (i.e., restricting sampling distributions to relevant subsets based on ancillary statistics) is somewhat elusive, but the idea is crystal clear in the conditional IM framework; see Theorem 2.

Remark 5. As we mentioned previously, conditional IMs and Theorem 1 have some connections to Bayes' theorem, in particular, in how information is combined or aggregated across samples. In fact, it can be shown that, in a certain sense, the Bayes solution is a special case of conditional IMs. To see this, consider a simple but generic example. The Bayes model, cast in terms of associations, is of the following form:

$$\theta = U_0, \quad U_0 \sim \mathsf{P}_{U_0} \quad \text{and} \quad X = a(U_0, U_1), \quad U_1 \sim \mathsf{P}_{U_1},$$

where P_U for $U = (U_0, U_1)$ is such that U_1 is conditionally independent given U_0. Here P_{U_0} is like the prior, and the distribution induced by $u_1 \mapsto a(\theta, u_1)$ given $U_0 = \theta$ determines the likelihood. It is clear that the function $a(U_0, U_1)$ is fully observed, so the conditional IM strategy would employ the conditional distribution of U_0 given the observed value x of $a(U_0, U_1)$. It is a simple exercise to see that the belief function in Theorem 1—the

one based on a "naive" IM—is exactly the Bayes posterior distribution function. So in any problem with a known prior distribution, the Bayes solution can be obtained as a special case of the conditional IM. No non-naive predictive random set is needed here because the naive IM itself is valid; this is consistent with the simple corresponding fact for posterior probabilities under a Bayes model with known prior.

Remark 6. As a follow-up to Remark 5, since a full prior is not required to construct a conditional IM, it is possible to develop an inferential framework based on conditional IMs and "partial prior information." For example, valid prior information may be available for some but not all components of θ. Incorporating the prior information where it is available while remaining prior-free where it is not can be obtained by slight extension of the argument in the previous remark. This important application of conditional IMs deserves further investigation.

3.4 Validity of conditional IMs

Here we extend the validity results in Martin and Liu (2012a) to the conditional IM context. The main obstacle is that the distribution function $\mathsf{P}_{\mathcal{S}}$, determined by the conditional distribution $\mathsf{P}_{V_T|H(x)}$ in Theorem 1, depends on data through the value $H(x)$. This is handled in Theorem 2 below by conditioning on the observed value of $H(X)$.

Fix $h \in H(\mathbb{X})$, and let $\mathbb{S}_h$ be a collection of $\mathsf{P}_{V_T|h}$-measurable subsets of $\mathbb{V}_T$. To keep notation simpler, assume that $\mathbb{S}_h$ contains both $\varnothing$ and $\mathbb{V}_T$. Like before, we also assume that $\mathbb{S}_h$ is nested in the sense that either $S \subseteq S'$ or $S' \subseteq S$ for all $S, S' \in \mathbb{S}_h$. Now we say that $\mathcal{S}$ is a *conditionally admissible* predictive random set, given h, if the support $\mathbb{S}_h$ is nested and if its distribution $\mathsf{P}_{\mathcal{S}|h}$ satisfies

$$\mathsf{P}_{\mathcal{S}|h}\{\mathcal{S} \subseteq K\} = \sup_{S \in \mathbb{S}_h : S \subseteq K} \mathsf{P}_{V_T|h}\{S\}, \quad K \subseteq \mathbb{V}_T. \tag{3.5}$$

So, in this case, the distribution of $\mathcal{S}$ depends on the particular h. With the help of Lemma 1 in Appendix A, we have the following extension of the validity theorem to the case of conditional IMs.

Theorem 2. *For any h, suppose that $\mathcal{S}$ is conditionally admissible, given h, with distribution $\mathsf{P}_{\mathcal{S}|h}$ as in* (3.5)*. If $\Theta_{T(x)}(\mathcal{S}) \neq \varnothing$ with $\mathsf{P}_{\mathcal{S}|h}$-probability 1 for all x such that $H(x) = h$, then the conditional IM is conditionally valid, i.e., for any $A \subseteq \Theta$,*

$$\sup_{\theta \notin A} \mathsf{P}_{X|\theta}\{\mathsf{bel}_{T(X)|h}(A; \mathcal{S}) \geq 1 - \alpha \mid H(X) = h\} \leq \alpha, \quad \forall\, \alpha \in (0,1). \tag{3.6}$$

Now is a good time to recall Remark 1. More general decompositions of the baseline association are allowed in Theorem 1, but only for the "separable" version (3.1a) is it possible to prove a conditional validity theorem. The point is that a condition like $c(X, U) = 0$ does not identify a fixed subset of the sample space on which probability calculations can be restricted—the subspace would depend on U.

Since the calibration property in Theorem 2 holds for all assertions A, we may translate (3.6) to a statement in terms of the corresponding plausibility function:

$$\sup_{\theta \in A} \mathsf{P}_{X|\theta}\{\mathsf{pl}_{T(X)|h}(A; \mathcal{S}) \leq \alpha \mid H(X) = h\} \leq \alpha, \quad \forall\, \alpha \in (0,1). \tag{3.7}$$

So, in addition to providing an objective scale for interpreting the conditional belief and plausibility function values, (3.7) provides desirable properties of conditional IM-based frequentist procedures. For example, if $h = H(x)$ is observed, the conditional $100(1-\alpha)\%$ plausibility region for θ is $\{\theta : \mathsf{pl}_{T(x)|h}(\theta; \mathcal{S}) > \alpha\}$. Then, by (3.7), the conditional coverage probability is $\mathsf{P}_{X|\theta}\{\mathsf{pl}_{T(X)|h}(\theta; \mathcal{S}) > \alpha \mid H(X) = h\} \geq 1 - \alpha$. In Fisher's mind, this is a more meaningful coverage probability since it is conditioned on a particular aspect of the observed data, namely, $H(x) = h$. In other words, the probability calculation focuses on a "relevant subset" $\{x : H(x) = h\}$ of the sample space. In some cases, though, conditional validity is the same as ordinary validity.

Corollary 1. *Suppose that the predictive random set $\mathcal{S}$ does not depend on the observed $H(x) = h$, so that $\mathsf{P}_{\mathcal{S}|h} \equiv \mathsf{P}_{\mathcal{S}}$ and $\mathsf{bel}_{T(x)|h} \equiv \mathsf{bel}_{T(x)}$. Then under the conditions of Theorem 2, the conditional IM is unconditionally valid, i.e., for any $A \subseteq \Theta$,*

$$\sup_{\theta \notin A} \mathsf{P}_{X|\theta}\{\mathsf{bel}_{T(X)}(A; \mathcal{S}) \geq 1 - \alpha\} \leq \alpha, \quad \forall\, \alpha \in (0, 1).$$

Two possible ways the condition of Corollary 1 may hold are as follows. First, in the P-step, the user may specify $\mathcal{S}$ directly without dependence on the observed $H(x) = h$; see Section 5.1. Second, it could happen that V_T and V_H are statistically independent, in which case the distribution $\mathsf{P}_{\mathcal{S}}$ for $\mathcal{S}$ is determined by the marginal distribution of V_T, which does not depend on h.

4 Finding conditional associations

4.1 Familiar things

In many problems, finding a decomposition (3.1) in Theorem 1 and the corresponding conditional association is easy to do. In general, the Neyman–Fisher factorization theorem implies we can define a conditional association through the marginal distribution of the minimal sufficient statistic $T(x)$. In standard problems, such as full-rank exponential families, the minimal sufficient statistics are easily obtained so this is probably the simplest approach. This, of course, includes both discrete and continuous problems. Similarly, if the problem has a group structure, invariance considerations can be used to find a decomposition; see Section 5.1. Note, however, that in light of Remark 3, one can consider other conditional associations if desirable. When the minimal sufficient statistic has dimension larger than that of the parameter, e.g., in curve exponential families, then some special conditioning is required; see Section 5.2.

4.2 A new differential equations-based technique

Here we describe a novel technique for finding conditional associations, based on differential equations. The method can be used for going directly from the baseline association to something lower-dimensional. In fact, in those nice problems mentioned above, it is easy to check that this differential equation-based technique reproduces the solutions based on minimal sufficiency, group invariance, etc. However, in our experience, this new approach is especially powerful in cases where the familiar things fail to give a fully satisfactory

reduction. In such cases, the differential equation-based technique can provide a further dimension reduction, beyond what sufficiency alone can give.

For concreteness, suppose $\Theta \subseteq \mathbb{R}$; the multi-parameter case can be handled similarly. The intuition is that ψ_T should map $\mathbb{U} \subseteq \mathbb{R}^n$ to Θ, so that $V_T = \psi_T(U)$ is one-dimensional, like θ. Moreover, ψ_H should map $\mathbb{U}$ into a $(n-1)$-dimensional manifold in $\mathbb{R}^n$, and be insensitive to changes in θ in the following sense. For baseline association $x = a(\theta, u)$, suppose that $u_{x,\theta}$ is the unique solution for u. Then for fixed x, we require that $\psi_H(u_{x,\theta})$ be constant in θ. In other words, we require that $\partial u_{x,\theta}/\partial\theta$ exists and

$$\underset{n\times 1}{0} = \frac{\partial \psi_H(u_{x,\theta})}{\partial \theta} = \underset{n \times n,\ \text{rank } n-1}{\frac{\partial \psi_H(u)}{\partial u}\Big|_{u=u_{x,\theta}}} \cdot \underset{n\times 1}{\frac{\partial u_{x,\theta}}{\partial \theta}}. \tag{4.1}$$

It is clear from the construction that, if a solution ψ_H of this (constrained) partial differential equation exists, then the value of $\psi_H(U)$ is fully observed, i.e., there is a corresponding function H, not depending on θ, such that $H(X) = \psi_2(U)$. So, with appropriate choice of ψ_T, the solution ψ_H of (4.1) determines the decomposition (3.1) in Theorem 1.

Formal theory on existence of solutions and on solving the differential equation system (4.1) is available. For example, the method of characteristics described in Polyanin et al. (2002) is powerful tool for solving such systems. However, such formalities here will take us too far off track. Examples of this method in action are given in Section 5.2, 6.4, and 6.5. In the first two cases, this differential equations method is applied after an initial step based on sufficiency, etc, provides an unsatisfactory dimension reduction.

5 Two detailed examples

5.1 A Student-t location problem

Suppose $X_1, \ldots, X_n$ is an independent sample from a Student-t distribution $\mathsf{t}_\nu(\theta)$, where the degrees of freedom ν is known but the location θ is unknown. This is a relatively challenging problem from a classical point of view because there is no satisfactory reduction via sufficiency. For the IM approach, start with a baseline association $X = \theta 1_n + U$, with $U = (U_1, \ldots, U_n)^\top$ and $U_i \sim \mathsf{t}_\nu$, independent, for $i = 1, \ldots, n$. For this location parameter problem, invariance considerations suggest the following decomposition:

$$X - T(X)1_n = U - T(U)1_n \quad \text{and} \quad T(X) = \theta + T(U),$$

where $T(\cdot)$ is the maximum likelihood estimator. Let $V_T = T(U)$ and $V_H = H(U) = U - T(U)1_n$. If h is the observed $H(X)$, then it follows from the result of Barndorff-Nielsen (1983) that the conditional distribution of V_T, given $V_H = h$, has a density

$$f_{\nu,h}(v_T) = c(\nu, h) \prod_{i=1}^n \bigl\{\nu + (v_T + h_i)^2\bigr\}^{-(\nu+1)/2},$$

where $c(\nu, h)$ is a normalizing constant that depends only on ν and h. If we write $F_{\nu,h}$ for the distribution function corresponding to the density $f_{\nu,h}$ above, then a conditional IM for θ can be built based on the following association (cf. Remark 2):

$$T(X) = \theta + F_{\nu,h}^{-1}(W), \quad W \sim \mathsf{Unif}(0,1).$$

		Coverage probability				Expected length			
		ν				ν			
Method	n	3	5	10	25	3	5	10	25
CIM	5	0.944	0.949	0.951	0.949	2.28	2.08	1.93	1.83
	10	0.949	0.951	0.952	0.953	1.56	1.45	1.35	1.29
	25	0.953	0.944	0.951	0.949	0.97	0.91	0.85	0.81
	50	0.953	0.951	0.953	0.947	0.68	0.64	0.60	0.58
MLE	5	0.915	0.933	0.944	0.940	2.10	1.98	1.88	1.81
	10	0.938	0.942	0.944	0.946	1.50	1.42	1.34	1.28
	25	0.946	0.939	0.943	0.940	0.96	0.90	0.85	0.81
	50	0.950	0.941	0.944	0.938	0.68	0.64	0.60	0.58

Table 1: Coverage probabilities and expected lengths of the 95% plausibility/confidence intervals for θ in the Student-t example based on, respectively, the conditional IM (CIM) and asymptotic normality of the maximum likelihood estimate (MLE).

With this conditional association, we are ready for the P- and C-steps. For simplicity, in the P-step we elect to take the predictive random set $\mathcal{S}$ as in (2.2); this also has some theoretical justification since $f_{\nu,h}$ should be approximately symmetric about $v_T = 0$ (Martin and Liu 2012a, Sec. 4.3.2). For the C-step, the random set $\Theta_{T(x)}(\mathcal{S})$ is

$$\big[T(x) - F_{\nu,h}^{-1}\big(0.5 + |W - 0.5|\big),\, T(x) - F_{\nu,h}^{-1}\big(0.5 - |W - 0.5|\big)\big], \quad W \sim \mathsf{Unif}(0,1).$$

From this point, numerical methods can be used to compute the conditional belief and plausibility functions. For example, if $A = \{\theta\}$ is a singleton assertion, then

$$\mathsf{pl}_{T(x)|h}(\theta; \mathcal{S}) = 1 - \big|1 - 2F_{\nu,h}(\theta - T(x))\big|,$$

and the corresponding $100(1-\alpha)\%$ plausibility interval for θ is

$$\{\theta : \mathsf{pl}_{T(x)|h}(\theta; \mathcal{S}) > \alpha\} = \big(T(x) + F_{\nu,h}^{-1}(\alpha/2),\, T(x) + F_{\nu,h}^{-1}(1-\alpha/2)\big).$$

For illustration, we present the results of a simple simulation study. In particular, for several pairs (n, ν), 5000 Monte Carlo samples of size n are obtained from a Student-t distribution with ν degrees of freedom and center $\theta = 0$. For each sample, the 95% plausibility interval for θ based on the conditional IM above is obtained. For comparison, we also compute the 95% confidence interval based on the asymptotic normality of the maximum likelihood estimate. The results of this simulation are summarized in Table 1. The general message is that while the classical confidence intervals are a bit shorter than the conditional IM plausibility intervals on average, the former tend to undershoot the target coverage probability while the latter are typically on target.

To conclude this example, recall our argument for conditional IM uniqueness in Remark 3. Here we can make a stronger statement. In the Student-t simulation example above, we also did the calculations with an alternative decomposition which took $V_T = U_1$ and $V_H = (0, U_2 - U_1, \ldots, U_n - U_1)$. Although Remark 3 suggests it, we were surprised to see that the results obtained with this "naive" decomposition were indistinguishable from those based on the arguably more reasonable maximum likelihood-driven decomposition. This suggests that the choice of decomposition does not affect the final results, provided that the conditioning is done correctly.

5.2 Fisher's problem of the Nile

Suppose two independent exponential samples, namely $X_1 = (X_{11}, \ldots, X_{1n})$ and $X_2 = (X_{21}, \ldots, X_{2n})$, are available, the first with mean θ^{-1} and the second with mean θ. The goal is to make inference on $\theta > 0$. The name comes from an application (Fisher 1973) to fertility of land in the Nile river valley. In this example, the maximum likelihood estimate is not sufficient, so conditioning on an ancillary statistic is recommended.

Sufficiency considerations suggest the following initial dimension reduction step:

$$S(X_1) = \theta^{-1} U_1 \quad \text{and} \quad S(X_2) = \theta U_2, \quad U_1, U_2 \sim \mathsf{Gamma}(n, 1),$$

where $S(X_i) = \sum_{j=1}^n X_{ij}$. But efficiency can be gained by considering a further reduction to a scalar auxiliary variable. Here we employ the differential equation technique in Section 4.2. Start with $u_{x,\theta} = (\theta S(x_1), \theta^{-1} S(x_2))^\top$. Differentiating with respect to θ reveals that our (real valued) conditioning function ψ_H must satisfy

$$\frac{\partial \psi_H(u)}{\partial u}\Big|_{u=u_{x,\theta}} \begin{pmatrix} S(x_1) \\ -\theta^{-2} S(x_2) \end{pmatrix} = 0.$$

If we take $\psi_H(u) = \{u_1 u_2\}^{1/2}$, then

$$\frac{\partial \psi_H(u)}{\partial u}\Big|_{u=u_{x,\theta}} = \frac{1}{2\{S(x_1)S(x_2)\}^{1/2}} \big(\theta^{-1} S(x_2),\, \theta S(x_1)\big)$$

and, clearly, this satisfies the differential equation above. Therefore, for (3.1), we take

$$H(X) = V_H \quad \text{and} \quad T(X) = \theta V_T, \tag{5.1}$$

where $T(X) = \{S(X_1)/S(X_2)\}^{1/2}$, $H(X) = \{S(X_1)S(X_2)\}^{1/2}$, $V_T = \{U_1/U_2\}^{1/2}$, and $V_H = \{U_1 U_2\}^{1/2}$. These quantities are familiar from the classical approach: $T(X)$ is the maximum likelihood estimate of θ, $H(X)$ is an ancillary statistic, and the pair $(T, H)(X)$ is a jointly minimal sufficient statistic (Ghosh et al. 2010).

By Theorem 1 and (5.1), we can focus on a conditional association based on $T(X) = \theta V_T$. The conditional distribution of V_T given $V_H = h$ is a generalized inverse Gaussian distribution (Barndorff-Nielsen 1977) with density function

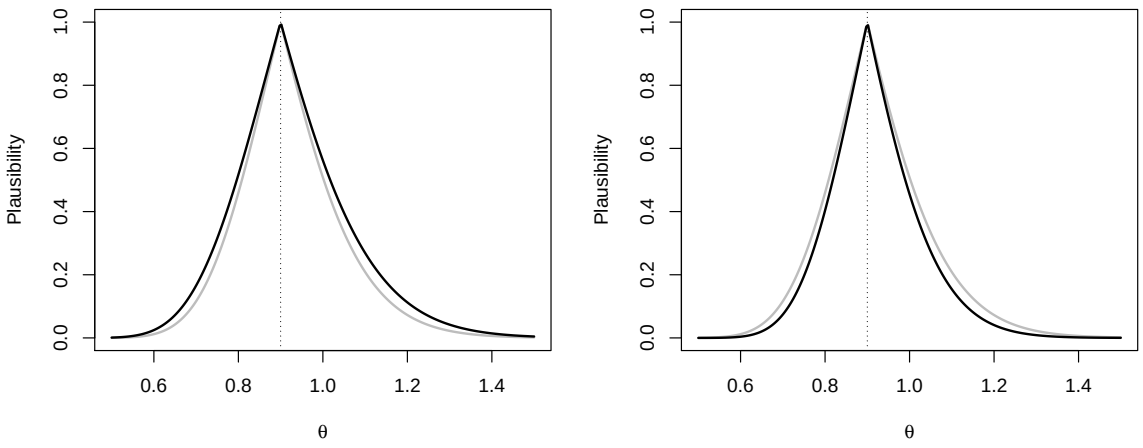
$$f_h(v_T) = \frac{1}{2 v_T K_0(2h)} \exp\{-h(v_T^{-1} + v_T)\}, \tag{5.2}$$

where K_0 is the modified Bessel function of the second kind. As a final simplifying step (cf. Remark 2), write the conditional association as

$$T(X) = \theta F_h^{-1}(W), \quad W \sim \mathsf{Unif}(0, 1), \tag{5.3}$$

where F_h is the distribution function corresponding to the density f_h in (5.2). This completes the A-step. If we take $\mathcal{S}$ as in (2.2) for the P-step, then the C-step gives

$$\Theta_{T(x)}(\mathcal{S}) = \Big[\frac{T(x)}{F_h^{-1}(0.5 + |W - 0.5|)},\, \frac{T(x)}{F_h^{-1}(0.5 - |W - 0.5|)}\Big], \quad W \sim \mathsf{Unif}(0, 1).$$



(a) $h = 15$ (b) $h = 25$

Figure 1: Plausibility functions for the conditional IM (black) and the "naive" conditional IM (gray) in the Nile example, with $T = 0.90$, $n = 20$, and the true $\theta = 1$. Gray curves in the two plots are the same since the naive conditional IM does not depend on h.

From this, the conditional belief/plausibility functions are readily evaluated.

For illustration, we display plausibility functions $\mathsf{pl}_t(\theta; \mathcal{S})$ for two conditional IMs. The first is based on that derived above; the second is based on a similar derivation, but we ignore V_H and simply work with the marginal distribution of V_T in (5.1). Figure 1 shows plausibility functions for $T(x) = 0.90$, with $n = 20$ and true $\theta = 1$, sampled from its conditional distribution given h, for two different values of h. In this case, if h is large (i.e., $h > n$), then the bona fide conditional IM has narrower level sets than the naive conditional IM. The opposite is true when h is small (i.e., $h < n$). This is due to the fact that the conditional Fisher information in T is an increasing function in h; see Ghosh et al. (2010, Example 1). Therefore, T has more variability when h is small, and this adjustment should be reflected in the plausibility function. The bona fide conditional IM catches this phenomenon while the naive one does not.

6 Local conditional IMs

6.1 Motivation

So far we have seen that the conditional IM approach is successful in problems where the baseline association admits a decomposition of the form (3.1a). However, as alluded to above, there are interesting and important problems where apparently no such decomposition exists. Next is one such problem, which may be considered as a "benchmark example" for conditional inference (Ghosh et al. 2010, Example 5).

Suppose $(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n})$ is an independent sample from a standard bivariate normal distribution with zero means, unit variances, but unknown correlation coefficient $\theta \in (-1, 1)$. A natural first step towards inference on θ is to take advantage of the

fact that $X_1 + X_2$ and $X_1 - X_2$ are independent. In particular, by defining

$$X_1 \leftarrow \frac{1}{2}\sum_{i=1}^{n}(X_{1i} + X_{2i})^2 \quad \text{and} \quad X_2 \leftarrow \frac{1}{2}\sum_{i=1}^{n}(X_{1i} - X_{2i})^2,$$

we may rewrite the baseline association as

$$X_1 = (1+\theta)U_1 \quad \text{and} \quad X_2 = (1-\theta)U_2, \quad U_1, U_2 \sim \mathsf{ChiSq}(n). \tag{6.1}$$

Sufficiency justifies this first reduction. Equation (6.1) is equivalent to

$$\frac{X_1}{U_1} + \frac{X_2}{U_2} = 2 \quad \text{and} \quad \frac{X_1}{X_2} = \frac{1+\theta}{1-\theta}\frac{U_1}{U_2}. \tag{6.2}$$

The first equation depends on data and auxiliary variable—free of θ—while the second depends also on θ. But note that the second expression in (6.2) is not of the form specified in (3.1a). In fact, this first expression is of the more general "non-separable" form $c(X, U) = 0$ described in Remark 1. So, although (6.2) provides a suitable decomposition of the baseline association, the requirements of Theorem 2 are not met, so the resulting conditional IM may not be valid.

6.2 Relaxing (3.1a) via localization

As describe above, the separability in (3.1a) can be too strict, but extending the conditional validity theorem to allow non-separablility appears difficult. The idea here is to relax (3.1a) in a different direction. Specifically, we propose to allow the pair of function (H, ψ_H) in (3.1a) to depend, locally, on the parameter. This generalization allows us some additional flexibility in finding an auxiliary variable dimension reduction.

Start by fixing an arbitrary $\theta_0 \in \Theta$. As in Theorem 1, consider a pair of function (T, H_{θ_0}), depending on θ_0, such that $x \mapsto (T(x), H_{\theta_0}(x))$ is one-to-one. Now take the corresponding functions $u \mapsto (\psi_T(u), \psi_{H,\theta_0}(u))$, one-to-one, such that the baseline association, at $\theta = \theta_0$, can be decomposed as

$$H_{\theta_0}(X) = \psi_{H,\theta_0}(U) \quad \text{and} \quad T(X) = a_T(\psi_T(U), \theta_0). \tag{6.3}$$

That is, (6.3), with $U \sim \mathsf{P}_U$, describes the sampling distribution $X \sim \mathsf{P}_{X|\theta_0}$. Suppose $H_{\theta_0}(X) = h_0$ is observed. We can compute the conditional distribution $\mathsf{P}_{V_T|h_0,\theta_0}$ of $V_T = \psi_T(U)$ given $\psi_{H,\theta_0}(U) = h_0$, which is then used to construct predictive random sets.

From this point, we may proceed exactly as before. That is, for the A-step, we get sets $\Theta_{T(x)}(v_T) = \{\theta : T(x) = a_T(v_T, \theta)\}$ just like before. For the P-step, we pick a conditionally admissible predictive random set $\mathcal{S} \sim \mathsf{P}_{S|h_0,\theta_0}$. Finally, the C-step produces conditional plausibility function

$$\mathsf{pl}_{T(x)|h_0,\theta_0}(A; \mathcal{S}) = 1 - \mathsf{P}_{\mathcal{S}|h_0,\theta_0}\{\Theta_{T(x)}(\mathcal{S}) \subseteq A^c\}, \quad A \subseteq \Theta.$$

We shall refer to the corresponding conditional IM as a *local* conditional IM at $\theta = \theta_0$. The adjective "local" is meant to indicate the dependence of the construction on the particular point θ_0. As we see below, the validity properties of this local conditional IM are, in a certain sense, also local.

6.3 Validity of local conditional IMs

The following theorem shows that for each θ_0 value, the local conditional IM at θ_0 is valid for some important assertions depending on the particular θ_0. The proof is exactly like that of Theorem 2 and, hence, omitted.

Theorem 3. *For any θ_0, take $h_0 \in H_{\theta_0}(\mathbb{X})$. Suppose that $\mathcal{S} \sim \mathsf{P}_{\mathcal{S}|h_0,\theta_0}$ is conditionally admissible. If $\Theta_{T(x)}(\mathcal{S}) \neq \varnothing$ with $\mathsf{P}_{\mathcal{S}|h_0,\theta_0}$-probability 1 for all x such that $H_{\theta_0}(x) = h_0$, then the local conditional IM at θ_0 is conditionally valid for $A = \{\theta_0\}$, i.e.,*

$$\mathsf{P}_{X|\theta_0}\{\mathsf{pl}_{T(X)|h_0,\theta_0}(\theta_0; \mathcal{S}) \leq \alpha \mid H_{\theta_0}(X) = h_0\} \leq \alpha, \quad \forall\, \alpha \in (0,1).$$

The validity result here is not as strong as in Theorem 2, a consequence of the localization. It does, however, imply that the local conditional plausibility region,

$$\{\theta : \mathsf{pl}_{T(x)|H_\theta(x)}(\theta; \mathcal{S}) > \alpha\}, \tag{6.4}$$

has the nominal (conditional) $1-\alpha$ coverage probability. This theoretical result is confirmed by the simulation experiment in Section 6.4 below. Observe that, in the definition of conditional plausibility region (6.4), the plausibility function depends on θ in two places—in the argument (the assertion) and in the local conditional IM itself. The latter structural dependence of the IM on the particular assertion is consistent with the optimality developments described in Martin and Liu (2012a).

6.4 Bivariate normal model, revisited

Here we demonstrate that the localization technique can be successfully used to solve the bivariate normal problem described above. Start with the relation in (6.1). Fix θ_0. To construct the functions (H, ψ_{H,θ_0}), depending on θ_0, and the corresponding local conditional IM at θ_0, we shall modify the differential equation approach in Section 4.2.

In this case, if we let $u_{x,\theta} = (x_1/(1+\theta), x_2/(1-\theta))^\top$, then we have

$$\frac{\partial u_{x,\theta}}{\partial \theta} = \Big(-\frac{x_1}{(1+\theta)^2}, \frac{x_2}{(1-\theta)^2}\Big)^\top.$$

For a local conditional IM at θ_0, we propose to choose a real-valued $\psi_{H,\theta_0}(u)$ such that $\partial \psi_{H,\theta_0}(u_{x,\theta})$ vanishes at $\theta = \theta_0$. If we take

$$\psi_{H,\theta_0}(u) = (1+\theta_0)\log u_1 + (1-\theta_0)\log u_2, \tag{6.5}$$

then

$$\frac{\partial \psi_{H,\theta_0}(u)}{\partial u} = \Big(\frac{1+\theta_0}{u_1}, \frac{1-\theta_0}{u_2}\Big),$$

so the derivative of $\psi_{H_0}(u_{x,\theta})$ with respect to θ is

$$\begin{aligned}
\frac{\partial \psi_{H,\theta_0}(u_{x,\theta})}{\partial \theta} &= \frac{\partial \psi_{H,\theta_0}(u)}{\partial u}\Big|_{u=u_{x,\theta}} \cdot \frac{\partial u_{x,\theta}}{\partial \theta} \\
&= -\frac{(1+\theta_0)^2}{x_1} \cdot \frac{x_1}{(1+\theta)^2} + \frac{(1-\theta_0)^2}{x_2} \cdot \frac{x_2}{(1-\theta)^2} \\
&= -\frac{(1+\theta_0)^2}{(1+\theta)^2} + \frac{(1-\theta_0)^2}{(1-\theta)^2}.
\end{aligned}$$

The latter expression clearly evaluates to zero at $\theta = \theta_0$, so ψ_{H,θ_0} satisfies the desired differential equation. The corresponding function $H(x) = H_{\theta_0}(x)$ is given by

$$H_{\theta_0}(x) = (1+\theta_0)\log\{x_1/(1+\theta_0)\} + (1-\theta_0)\log\{x_2/(1-\theta_0)\}.$$

For the local conditional association—the second expression in (6.3)—we take

$$T(X) = z(\theta) + V_T,$$

where $T(x) = \log(x_1/x_2)$, $z(\theta) = \log\{(1+\theta)/(1-\theta)\}$, and $V_T = T(U)$. Then $\mathsf{P}_{V_T|\theta_0,h_0}$ is the conditional distribution of V_T, given (θ_0, h_0), where h_0 is the observed $H_{\theta_0}(X) = H_{\theta_0}(x)$. This conditional distribution has a density, given by

$$f_{h_0,\theta_0}(v_T) \propto \exp\big\{-n\theta_0 v_T/2 - \cosh(v_T/2)e^{(h_0-\theta_0 v_T)/2}\big\}.$$

If we let F_{h_0,θ_0} denote the corresponding distribution function, then (cf. Remark 2) we can describe this conditional association model by

$$T(X) = z(\theta) + F^{-1}_{h_0,\theta_0}(W), \quad W \sim \mathsf{Unif}(0,1).$$

If, for the P-step, we use the predictive random set $\mathcal{S}$ in (2.2), then the local conditional plausibility function is

$$\mathsf{pl}_{T(x)|h_0,\theta_0}(\theta_0;\mathcal{S}) = 1 - \big|1 - 2F_{h_0,\theta_0}(T(x) - z(\theta_0))\big|.$$

A local conditional $100(1-\alpha)\%$ plausibility interval for θ can be found just as before, by thresholding the plausibility function at α. It follows from Theorem 3 that these intervals will have the nominal coverage probabilities.

For illustration, we consider a simple simulation experiment. In particular, we compute the local conditional 95% plausibility interval for θ in for 5000 Monte Carlo samples based on $\theta = 0.3$. For several values of n, the estimated coverage probabilities and expected lengths are compared, in Table 2, to those of the conditional frequentist interval based on the so-called "$r^\star$" approximation due to Barndorff-Nielsen (1986) and Fraser (1990), summarized nicely in Reid (1995, 2003). As in Section 5.1, the general message is that the conditional frequentist interval is a bit shorter than the generalized conditional IM plausibility interval on average, but the former tends to undershoot the target coverage probability 0.95 while the latter is on target for all n. That the frequentist intervals are conditioned on the observed value of an ancillary statistic makes them easier to interpret compared to an unconditional interval, but they still lack the probabilistic interpretation of the plausibility intervals.

6.5 A variance-components example

Suppose $X_1, \ldots, X_n$ are independent with $X_i \sim \mathsf{N}(0, 1+w_i\theta)$, $i = 1, \ldots, n$, where the w_i's are known constants (not all equal) and inference on the unknown $\theta \geq 0$ is is desired. Such a model arises in an unbalanced hierarchical analysis of variance model framework: given μ_i, $Y_{i1}, \ldots, Y_{iw_i}$ are independent samples from $\mathsf{N}(\mu_i, 1)$, $i = 1, \ldots, n$, and $\mu_1, \ldots, \mu_n$ are independent $\mathsf{N}(0, \theta)$. In this case, w_i is the size of the ith group, $i = 1, \ldots, n$. Set

	Coverage probability		Expected length	
n	LCIM	$r^\star$	LCIM	$r^\star$
10	0.951	0.912	0.961	0.888
25	0.952	0.935	0.663	0.621
50	0.946	0.936	0.480	0.455
100	0.952	0.943	0.341	0.328
1000	0.947	0.938	0.108	0.104
10000	0.952	0.943	0.034	0.033

Table 2: Coverage probabilities and expected lengths of 95% plausibility and confidence intervals for the correlation θ in the bivariate normal problem based on, respectively, the local conditional IM (LCIM) and the $r^\star$ approach reviewed by Reid (2003).

$X_i = w_i^{1/2}\bar{Y}_{i\cdot}$, a rescaled version of the ith group mean. Then the marginal distribution of X_i is $\mathsf{N}(0, 1 + w_i\theta)$, the one assumed here.

Although the model is relatively simple, inference on θ remains a challenge, partly due to the fact that the primary question of interest is whether $\theta = 0$, a boundary point. In particular, the case $\theta = 0$ corresponds to all the μ_i's in the hierarchical framework being equal, i.e., none of the "treatments" are significant. Here we develop a local conditional IM approach for inference on θ, with focus on testing $H_0 : \theta = 0$ versus the full one-sided alternative $H_1 : \theta > 0$. Confidence and Bayesian credible intervals for θ are developed in Zhang and Woodroofe (2002); see, also, Martin (2012). These latter papers focus on a balanced but more general version where the mean of μ_i's and/or the variances σ_i^2 of the Y_{ij}'s are unknown but the "weights" w_i are all equal. These more general problems can be considered in the IM framework, but a special marginalization technique is required which is beyond our present scope; see Martin and Liu (2012b).

Consider the following baseline association:

$$X_i^2 = (1 + w_i\theta)U_i, \quad U_i \sim \mathsf{ChiSq}(1), \quad i = 1, \ldots, n. \tag{6.6}$$

To see that a local conditional IM is required, take, for the moment, the special case $n = 2$. Then a θ-free function of observable data and unobservable auxiliary variables would be something like

$$\frac{X_2^2}{U_2} - \frac{w_2}{w_1}\frac{X_1^2}{U_1} = 1 - \frac{w_2}{w_1}.$$

But the left-hand side of this display is clearly not of the separable form (3.1a), just as in the bivariate normal example described in Section 6.1 above. Therefore, a regular conditional IM may not be valid so we should look for a local conditional IM.

To find a local conditional IM, we shall employ the PDE technique presented above. Start by writing the relationship (6.6) in terms of u for given x, θ:

$$u_{x,\theta,i} = \frac{x_i^2}{1 + w_i\theta}, \quad i = 1, \ldots, n.$$

The derivative of this quantity with respect to θ for fixed x is

$$\frac{\partial u_{x,\theta,i}}{\partial\theta} = -\frac{w_i x_i^2}{(1 + w_i\theta)^2} = -\frac{w_i}{1 + w_i\theta}u_{x,\theta,i}, \quad i = 1, \ldots, n.$$

The method of characteristics (Polyanin et al. 2002) for solving PDEs identifies the function $\psi_{H,\theta}(\cdot)$ from $\mathbb{R}^n$ to $\mathbb{R}^{n-1}$, given by

$$\psi_{H,\theta}(u)_i = c_{i+1}(\theta)\log u_{i+1} - c_i(\theta)\log u_i, \quad i = 1,\ldots,n-1,$$

where $c_i(\theta) = (1+w_i\theta)/w_i$. The derivative of this function with respect to u is a $(n-1)\times n$ matrix, whose ith row looks like

$$\Big(0,\ldots,0,-\frac{c_i(\theta)}{u_i},\frac{c_{i+1}(\theta)}{u_{i+1}},0,\ldots,0\Big), \quad i = 1,\ldots,n-1. \tag{6.7}$$

It is a simple calculation to check that, for each $i = 1,\ldots,n-1$, the row vector above, evaluated at $u = u_{x,\theta}$, is orthogonal to the vector $\partial u_{x,\theta}/\partial\theta$. Therefore, if we fix $\theta = \theta_0$, then $\psi_{H,\theta_0}(u)$ is indeed a solution to the PDE

$$0 = \frac{\partial\psi_{H,\theta_0}(u_{x,\theta})}{\partial\theta} = \frac{\partial\psi_{H,\theta_0}(u)}{\partial u}\Big|_{u=u_{x,\theta}}\frac{\partial u_{x,\theta}}{\partial\theta}, \quad \text{at } \theta = \theta_0.$$

We then have a decomposition (6.3) of the baseline association (6.6) given by

$$\sum_{i=1}^n \log X_i^2 = \sum_{i=1}^n \log(1+w_i\theta) + \sum_{i=1}^n \log U_i$$

for the $T(X) = a_T(\psi_T(U),\theta)$ part, and

$$c_{i+1}(\theta_0)\log\frac{X_{i+1}^2}{w_{i+1}c_{i+1}(\theta_0)} - c_i(\theta_0)\log\frac{X_i^2}{w_ic_i(\theta_0)} = \psi_{H,\theta_0}(U)_i, \quad i = 1,\ldots,n-1$$

for the $H_{\theta_0}(X) = \psi_{H,\theta_0}(U)$ part. Let $V_T = \sum_{i=1}^n \log U_i$. Then the conditional association can be written as

$$T(X) = z(\theta) + V_T, \quad V_T \sim \mathsf{P}_{V_T|\theta_0,h_0}$$

where $T(X) = \sum_{i=1}^n \log X_i^2$ and $\mathsf{P}_{V_T|\theta_0,h_0}$ is the conditional distribution of $V_T = \psi_T(U)$ given θ_0 and the observed value h_0 of $\psi_{H,\theta_0}(U)$. Since the relationship between U and $(\psi_T(U),\psi_{H,\theta_0}(U))$ is log-linear, i.e.,

$$\begin{pmatrix}\psi_{H,\theta_0}(U)_1\\ \vdots\\ \psi_{H,\theta_0}(U)_{n-1}\\ \psi_T(U)\end{pmatrix} = \begin{pmatrix}c_1\\ \vdots\\ c_{n-1}\\ 1_n^\top\end{pmatrix}\begin{pmatrix}\log U_1\\ \vdots\\ \log U_{n-1}\\ \log U_n\end{pmatrix},$$

where the $1\times n$ vectors $c_1,\ldots,c_{n-1}$ are given in (6.7), and the nth row is all 1's, one can easily find (numerically) the joint density and, hence, the conditional density $f_{\theta_0,h_0}(v_T)$ of V_T given $\psi_{H,\theta_0}(U) = h_0$. We omit the details of this calculation here. Therefore, the conditional association can be rewritten as

$$T(X) = z(\theta) + F_{\theta_0,h_0}^{-1}(R), \quad R \sim \mathsf{Unif}(0,1),$$

where F_{θ_0,h_0} is the distribution function corresponding to f_{θ_0,h_0}.

This looks similar to the conditional association in the bivariate normal example. However, for certain assertions of interest, there is an additional obstacle to be overcome in this case, namely, "conflict cases." That is, $T(X)$ and $F^{-1}_{\theta_0,h_0}(R)$ can be arbitrary real numbers, but $z(\theta)$ is non-negative, so not all pairs of (X, R) are compatible with the conditional association stated above. In other words, there is a special, almost imperceptible constraint which we must deal with. For problems with conflict cases, there is an efficient modification of the IM approach laid out in Ermini Leaf and Liu (2012); fortunately, for the problem of interest here, we will not need this modification.

In our case, we are interested in the assertion $A = \{0\}$, i.e., that the treatments are not significant. Since $\Theta = [0, \infty)$, A^c is a "one-sided" assertion and the arguments in Martin and Liu (2012a, Theorem 4) suggest that the optimal predictive random set in this problem is $\mathcal{S} = [0, R]$, $R \sim \mathsf{Unif}(0, 1)$. In this case, we take $\theta_0 = 0$ and we have

$$\Theta_t(r) = \{\theta : z(\theta) = t - F^{-1}_{0,h_0}(r)\}, \quad t \in \mathbb{R}, \quad r \in [0, 1].$$

With the optimal predictive random set $\mathcal{S}$, we then get

$$\Theta_t(\mathcal{S}) = \bigcup_{r \in \mathcal{S}} \Theta_t(r) = \{\theta : z(\theta) \geq t - F^{-1}_{0,h_0}(R)\}, \quad R \sim \mathsf{Unif}(0, 1).$$

These random sets are non-empty with $\mathsf{P}_{\mathcal{S}}$-probability 1, so we can effectively ignore the constraint mentioned above. But, if interest were in more general singleton assertions, say, for constructing a plausibility interval, then more care would be needed.

With this construction, it is easy to check that the plausibility function at $\theta = 0$ is $\mathsf{pl}_t(0; \mathcal{S}) = 1 - F_{0,h_0}(t)$, which can be evaluated numerically. A size-0.05 test of $H_0 : \theta = 0$ can be performed by rejecting H_0 if $\mathsf{pl}_t(0; \mathcal{S}) \leq 0.05$. A simulation study was performed to compare the power of this local conditional IM test with that of the parametric bootstrap likelihood ratio test. In our experiments, we found that the two tests had indistinguishable power functions. We believe this is a good sign. Recall that the local conditional IM is sacrificing something by focusing on validity only locally. But here we find that by choosing that particular point as the point of interest—in this case $\theta_0 = 0$—we can maintain the expected strong performance of the conditional IM. Indeed, the parametric bootstrap likelihood ratio test is exact and arguably a gold-standard for efficiency. So the fact that the local conditional IM can match this gold-standard suggests that nothing is lost by focusing on a suitably chosen point θ_0.

7 Discussion

In this paper we have extended the basic IM framework laid out in Martin and Liu (2012a) by developing an auxiliary variable dimension reduction strategy. This reduction simultaneous accomplishes two goals. First, it provides a suitable combination of information across samples, and we argue in Remarks 4 and 5 in Section 3.3 that Fisher's concept of sufficiency and Bayes' theorem can both be viewed as special cases of this combination of information via conditioning. Second, this reduction makes construction of efficient predictive random sets considerably simpler. An apparently new differential equation technique is proposed by which an auxiliary variable dimension reduction can

be found even in cases where traditional sufficiency considerations fail to give a satisfactory solution. In addition, as our simulation results in Sections 5.1 and 6.4 demonstrate, even with a default choice of predictive random set, the conditional IM results are as good or better than those using standard likelihood-based methods. This suggests that our proposed method of combining information is, in some sense, efficient. We expect that the conditional IM approach, paired with the optimal predictive random sets, will have even better performance. However, more work is needed since computation of these optimal predictive random sets remains a non-trivial task.

The local conditional IMs considered in Section 6 are an important contribution. Indeed, these tools provide a means to reduce the effective dimension even in cases where the minimal sufficient statistic has dimension greater than that of the parameter. For example, in the variance-components problem in Section 6.5, we identified a one-dimensional auxiliary variable to predict, even though there is no dimension reduction that can be achieved via sufficiency. The idea of focusing on validity locally at a single $\theta = \theta_0$ itself seems to provide an improvement, this is, in fact, a special case of a more general idea. One could measure locality by a general assertion A, not necessarily a singleton $A = \{\theta_0\}$. In this way, one can develop a conditional IM that focuses on validity at a particular assertion A, thus extending the range of application of local conditional IMs. Though a clear picture of this general idea is not yet available, it is certainly within reach.

The examples in this paper have focused on continuous distributions. Efficient inference in discrete problems is challenging in any framework, and IMs are no different. For nice discrete problems, e.g., regular exponential families, the IM analysis described herein can be carried out without a hitch. However, when sufficiency consideration alone provide inadequate auxiliary variable dimension reduction, new tools are needed. In particular, the differential equation-based technique used above may not be applicable because the baseline association is based on inequalities rather than equalities. But perhaps by using discrete auxiliary variables, as opposed to continuous ones, it may be possible to refine this differential equation-driven technique for application in discrete data problems. Further investigation along these lines is needed.

Finally, note that the problem considered here is when there is some sort of replication or information about a single quantity coming from multiple sources, e.g., several independent (noisy) measurements on the same quantity. In such cases, the goal is to combine the information coming from these different sources, and conditioning is shown to be the right tool for this sort of dimension reduction. In other problems, dimension reduction is needed because the real quantity of interest is some lower-dimensional characteristic of the full unknown parameter. For these nuisance parameter problems, a different sort of dimension reduction is needed, and marginalization is the appropriate tool. The companion paper (Martin and Liu 2012b) deals with this problem from an IM point of view, i.e., with a focus on efficient prediction of unobservable auxiliary variables.

Acknowledgments

This work is partially supported by the U.S. National Science Foundation, grants DMS-1007678, DMS-1208833, and DMS-1208841. The authors thank Dr. Jing-Shiang Hwang for helpful comments on an earlier draft.

A Proofs, etc

Proof of Theorem 1. For given $u \in \mathbb{U}$, let $(v_T, v_H) = (\psi_T(u), \psi_H(u))$. Let $\Theta_x(u) = \{\theta : x = a(u, \theta)\}$ be as defined in Section 2.1, and define

$$\Theta_{T(x)}(v_T) = \{\theta : T(x) = a_T(v_T, \theta)\}. \tag{A.1}$$

Pick any $A \subseteq \Theta$. It is clear that

$$\begin{aligned} \Theta_x(u) = \varnothing &\iff \Theta_{T(x)}(v_T) = \varnothing \text{ or } v_H \neq H(x), \\ \Theta_x(u) \subseteq A &\iff \Theta_{T(x)}(v_T) \subseteq A. \end{aligned}$$

By definition of conditional probability, and the assumed existence of the conditional distribution $\mathsf{P}_{V_T|H(x)}$ for each x, the belief function $\mathsf{bel}_x(A)$ for the baseline association with naive predictive random set $\mathcal{S} = \{U\}$, $U \sim \mathsf{P}_U$, can be re-expressed as

$$\begin{aligned} \mathsf{bel}_x(A) &= \mathsf{P}_U\{\Theta_x(U) \subseteq A \mid \Theta_x(U) \neq \varnothing\} \\ &= \mathsf{P}_{(V_T,V_H)}\{\Theta_{T(x)}(V_T) \subseteq A \mid \Theta_{T(x)}(V_T) \neq \varnothing,\, V_H = H(x)\} \\ &= \mathsf{P}_{V_T|H(x)}\{\Theta_{T(x)}(V_T) \subseteq A \mid \Theta_{T(x)}(V_T) \neq \varnothing\}, \end{aligned}$$

the latter quantity being the belief function for the naive IM from $(T(x), a_T, \mathsf{P}_{V_T|H(x)})$ with predictive random set $\mathcal{S} = \{V_T\}$, $V_T \sim \mathsf{P}_{V_T|H(x)}$. Since this holds for all x and all $A \subseteq \Theta$, the claimed equivalence follows. □

Lemma 1. *Fix $h \in H(\mathbb{X})$ and take $\mathcal{S}$ with natural measure $\mathsf{P}_{\mathcal{S}|h}$ as in Section 3.4. Write $Q_{\mathcal{S}|h}(v_T) = \mathsf{P}_{\mathcal{S}|h}\{\mathcal{S} \not\ni v_T\}$. Then, for all h, $Q_{\mathcal{S}|h}(V_T)$ is stochastically no larger than $\mathsf{Unif}(0,1)$ for $V_T \sim \mathsf{P}_{V_T|h}$.*

Proof. The goal is to show that $\mathsf{P}_{V_T|h}\{Q_{\mathcal{S}|h}(V_T) > 1-\alpha\} \leq \alpha$, for all $\alpha \in (0,1)$. Take any such α and set $S_\alpha = \bigcap\{S \in \mathbb{S}_h : \mathsf{P}_{V_T|h}(S) \geq 1-\alpha\}$. Since $\mathbb{S}_h$ is nested, it follows that $S_\alpha \in \mathbb{S}_h$ and $\mathsf{P}_{V_T|h}(S_\alpha) \geq 1-\alpha$. By (3.5), $\mathsf{P}_{\mathcal{S}|h}\{\mathcal{S} \subseteq S_\alpha\} \geq 1-\alpha$. Since $Q_{\mathcal{S}|h}(v_T) > 1-\alpha$ iff $v_T \notin S_\alpha$, it follows that

$$\mathsf{P}_{V_T|h}\{Q_{\mathcal{S}|h}(V_T) > 1-\alpha\} = \mathsf{P}_{V_T|h}(S_\alpha^c) = 1 - \mathsf{P}_{V_T|h}(S_\alpha) \leq \alpha.$$

The claim follows since h and α were arbitrary. □

Proof of Theorem 2. Take any $\theta \notin A$ as the true value of the parameter; then $T(X) = a_T(V_T, \theta)$, with $V_T \sim \mathsf{P}_{V_T|h}$, characterizes the conditional distribution of X, given $H(X) = h$. Since $A \subset \{\theta\}^c$, monotonicity of the belief function gives

$$\mathsf{bel}_{T(X)|h}(A; \mathcal{S}) \leq \mathsf{bel}_{T(X)|h}(\{\theta\}^c; \mathcal{S}) = \mathsf{P}_{\mathcal{S}|h}\{\Theta_{T(X)}(\mathcal{S}) \not\ni \theta\} = Q_{\mathcal{S}|h}(V_T).$$

Conditional admissibility of $\mathcal{S}$ implies that the right-hand side is stochastically no larger than $\mathsf{Unif}(0,1)$. This, in turn, implies the same of the left-hand side $\mathsf{bel}_{T(X)|h}(A; \mathcal{S})$, as a function of $X \sim \mathsf{P}_{X|\theta}$, given $H(X) = h$. Therefore,

$$\mathsf{P}_{X|\theta}\{\mathsf{bel}_{T(X)|h}(A; \mathcal{S}) \geq 1-\alpha \mid H(X) = h\} \leq \mathsf{P}\{\mathsf{Unif}(0,1) \geq 1-\alpha\} = \alpha.$$

Taking supremum over $\theta \notin A$ proves (3.6). □

Proof of Corollary 1. Since the distribution of $\mathcal{S}$ is free of h in this case, the belief function $\mathsf{bel}_{T(X)|h} \equiv \mathsf{bel}_{T(X)}$ is also free of h. Therefore, before taking supremum in the last line of the proof of Theorem 2, we can take expectation over h to remove the conditioning, so that the validity property holds unconditionally, like in (2.6). □

References

Barndorff-Nielsen, O. (1977), "Exponentially decreasing distributions for the logarithm of particle size," *Proc. R. Soc. Lond. A.*, 353, 401–419.

— (1983), "On a formula for the distribution of the maximum likelihood estimator," *Biometrika*, 70, 343–365.

Barndorff-Nielsen, O. E. (1986), "Inference on full or partial parameters based on the standardized signed log likelihood ratio," *Biometrika*, 73, 307–322.

Berger, J. (2006), "The case for objective Bayesian analysis," *Bayesian Anal.*, 1, 385–402.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009), "The formal definition of reference priors," *Ann. Statist.*, 37, 905–938.

Cox, D. R. (1980), "Local ancillarity," *Biometrika*, 67, 279–286.

Dempster, A. P. (2008), "Dempster–Shafer calculus for statisticians," *Internat. J. of Approx. Reason.*, 48, 265–277.

Ermini Leaf, D. and Liu, C. (2012), "Inference about constrained parameters using the elastic belief method," *Internat. J. Approx. Reason.*, 53, 709–727.

Fisher, R. A. (1973), *Statistical methods and scientific inference*, New York: Hafner Press, 3rd ed.

Fraser, D. A. S. (1990), "Tail probabilities from observed likelihoods," *Biometrika*, 77, 65–76.

— (2011), "Is Bayes posterior just quick and dirty confidence?" *Statist. Sci.*, 26, 299–316.

Fraser, D. A. S., Reid, N., Marras, E., and Yi, G. Y. (2010), "Default priors for Bayesian and frequentist inference," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72, 631–654.

Ghosh, M., Reid, N., and Fraser, D. A. S. (2010), "Ancillary statistics: a review," *Statist. Sinica*, 20, 1309–1332.

Hannig, J. (2009), "On generalized fiducial inference," *Statist. Sinica*, 19, 491–544.

— (2012), "Generalized fiducial inference via discretization," *Statist. Sinica*, to appear.

Hannig, J. and Lee, T. C. M. (2009), "Generalized fiducial inference for wavelet regression," *Biometrika*, 96, 847–860.

Martin, R. (2012), "Plausibility functions and exact frequentist inference," Unpublished manuscript, `arXiv:1203.6665`.

Martin, R., Ermini Leaf, D., and Liu, C. (2012), "Optimal inferential models for a Poisson mean," Unpublished manuscript, `arXiv:1207.0105`.

Martin, R. and Liu, C. (2012a), "Inferential models: A framework for prior-free posterior probabilistic inference," *J. Amer. Statist. Assoc.*, to appear, `arXiv:1206.4091`.

— (2012b), "Marginal inferential models: redirecting information for prior-free probabilistic inference," Unpublished manuscript.

Martin, R., Zhang, J., and Liu, C. (2010), "Dempster–Shafer theory and statistical inference with weak beliefs," *Statist. Sci.*, 25, 72–87.

Polyanin, A. D., Zaitsev, V. F., and Moussiaux, A. (2002), *Handbook of first order partial differential equations*, vol. 1 of *Differential and Integral Equations and Their Applications*, London: Taylor & Francis Ltd.

Reid, N. (1995), "The roles of conditioning in inference," *Statist. Sci.*, 10, 138–157.

— (2003), "Asymptotics and the theory of inference," *Ann. Statist.*, 31, 1695–1731.

Shafer, G. (2011), "A betting interpretation for probabilities and Dempster-Shafer degrees of belief," *Internat. J. Approx. Reason.*, 52, 127–136.

Xie, M. and Singh, K. (2012), "Confidence distribution, the frequentist distribution of a parameter – a review," *Int. Statist. Rev.*, to appear.

Xie, M., Singh, K., and Strawderman, W. E. (2011), "Confidence distributions and a unifying framework for meta-analysis," *J. Amer. Statist. Assoc.*, 106, 320–333.

Zhang, J. and Liu, C. (2011), "Dempster–Shafer inference with weak beliefs," *Statist. Sinica*, 21, 475–494.

Zhang, T. and Woodroofe, M. (2002), "Credible and confidence sets for the ratio of variance components in the balanced one-way model," *Sankhyā Ser. A*, 64, 545–560, special issue in memory of D. Basu.